Problem 1. (Naive Bayes)
In this problem we aim to predict the possibility of failure during manufacturing an item based on the pressure and temperature applied to the object. Consider the following hypothetical dataset consisting of 10 measurements:

temperature, $^\circ F$	281	126	111	244	288	428	124	470	301	323
pressure, MPa	262	125	282	226	119	155	209	291	292	281
failure	1	0	0	0	0	1	0	1	0	1

1. Given each of the two classes 1 (failure) and 0 (no failure), determine the mean and the variance for each feature vector. Based on this, determine the corresponding 2 Gaussian distributions for each of the two features.

2. Use the Gaussian Naive Bayes approach to determine whether the test point $x^{\text{test}} = (270, 170)$ would lead to a failure or not.

Problem 2. (kNN)
In class, we discussed that the kNN approach can be used for both regression and classification. Consider a regression problem with a dataset $\{x^i, y^i\}_{i=1}^N$, $x_i \in \mathbb{R}^d$, $y^i \in \mathbb{R}$.

1. Explain why you need to normalize the data. Write the equation for data normalization based on the so-called Z-score.

2. Given a test point x, write the formula for determining the estimated label $\hat{y} \in \mathbb{R}$ using $K = 1$ nearest neighbor, and 2-norm distance (referred to as Euclidean distance) as well as the 1-norm distance (referred to as Manhattan distance).

3. Repeat the same task as with $K = 2$ nearest neighbors. Hint: for regression, we take the label corresponding to the average of the K neighbor labels.

4. How would you determine which distance metric and which K to use?

Problem 3. (Naive Bayes, k-Nearest Neighbor, taken from Exam 2023)
Company X is hiring employees and aims to hire those who spend less time watching videos online. Thus, it wants to predict an applicant's potential to watch online videos based on past employee data. In particular, for each of the 1000 past employees, it has recorded whether they have a high GPA, and whether they watch online videos. Among those who do not watch videos, some play sports (none of those who watch videos play sports). The survey is summarized below.

	Employees	Sport	No Video
High GPA	600	150	300
low GPA	400	50	100

Let event A denote having a high GPA, event B denote playing sports, and event C denote not watching videos.

1. Calculate the following: (a) empirical probability of having a high GPA; (b) empirical probability of having a high GPA and playing sports.

2. Show that the events A and B are not independent.

3. Calculate the following: (a) the empirical conditional probability of event A given event C; (b) the empirical conditional probability of events A and B given event C.

4. Show that conditioned on C, the events A and B are independent.

The company aims to have a classifier for an employee by using the exact GPA ($x_1 \in \mathbb{R}_+$) and the average number of hours of sport played per week ($x_2 \in \mathbb{R}_+$). Based on past data, it fits two probability density functions conditioned on Y, where Y corresponds to watching videos ($Y = 1$) or not ($Y = 0$). These are denoted by $f_{x_1|Y} : \mathbb{R} \to \mathbb{R}$ and $f_{x_2|Y} : \mathbb{R} \to \mathbb{R}$.

5. Formulate the Naive Bayes classifier.

6. For an applicant, the company has obtained x_1, x_2 from which it has evaluated $f_{x_1|0}(x_1) = 0.25$, $f_{x_2|0}(x_2) = 2.00$, $f_{x_1|1}(x_1) = 0.20$ and $f_{x_2|1}(x_2) = 2.20$. Based on the Naive Bayes classifier, would this person be likely to watch online videos at work?

7. On a test set obtained from recently hired employees, it was found that the classifier has more false positives than false negatives. The company decided to change the prior on the probability of watching videos (perhaps the new generation has been bored of all the online videos). What term(s) in your Naive Bayes classifier would you change and how?

Problem 4. (Neural networks function class expressiveness)
Consider a data set $\{x^i, y^i\}_{i=1}^N$, with $x^i \in \mathbb{R}^d$ and $y^i \in \mathbb{R}^m$. Let our predictor be a neural network $f : \mathbb{R}^d \to \mathbb{R}^m$.

1. Consider a neural network with one hidden layer having 3 nodes and an output layer having m nodes, with activation function $g : \mathbb{R} \to \mathbb{R}$ for each node. How many parameters need to be determined in this network?

2. Recall the definition of an affine function from your "Background and notations.pdf" file (posted on Moodle, 9 September - 15 September). Show that if the activation function for each node in the hidden layer and each node in the output layer is the identity, $g(z) = z$, $\forall z \in \mathbb{R}$, then the neural network predictor is the same as a linear predictor with $m(d + 1)$ parameters to be determined. Hence, the problem is the same as linear regression.

3. For the case of linear regression above, consider using batches of size N_b for training the network. Consider the square-error loss function. Write the pseudo-code for stochastic gradient descent with this batch size. For simplicity, let $m = 1$.

Problem 5. (Neural networks expressive power)
Consider a neural network

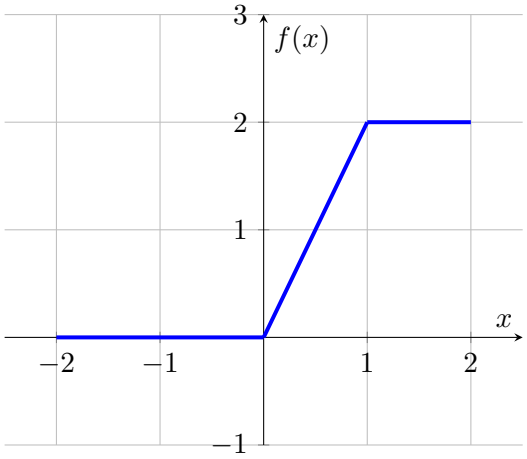
$$f : [-2, 2] \to \mathbb{R}, \quad f(x) = W^{[1]T} g\left(W^{[0]T} x + b^{[0]}\right) + b^{[1]},$$

with a single hidden layer and the ReLU activation function[1] $g(x) := \max(0, x)$. Supposing that there are two nodes in the hidden layer, determine the weight matrices $W^{[0]} \in \mathbb{R}^{1\times 2}, W^{[1]} \in \mathbb{R}^{2\times 1}$ and the biases $b^{[0]} \in \mathbb{R}^2, b^{[1]} \in \mathbb{R}$ such that:

1. $f(x) = x$.

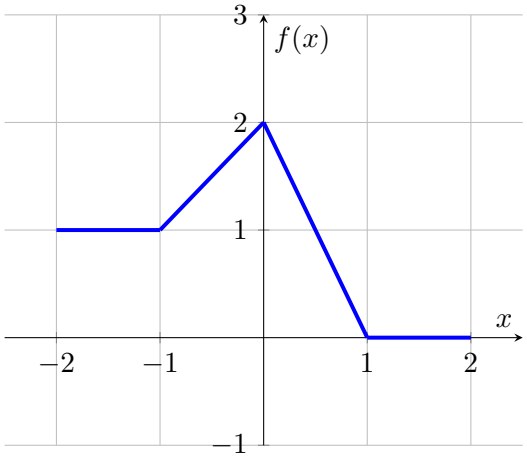
[1] Activation functions are applied elementwise to each node of a hidden layer.

2. f has the following graph:



Supposing that there are three nodes in the hidden layer, determine the weights $W^{[0]} \in \mathbb{R}^{1\times 3}$, $W^{[1]} \in \mathbb{R}^{3\times 1}$ and the biases $b^{[0]} \in \mathbb{R}^3, b^{[1]} \in \mathbb{R}$ such that:

3. f has the following graph:



Problem 6. (Convolutional neural networks from the python exercise)
Consider the convolutional neural network exercise with the MNIST dataset.

1. Write the dimensionals of the input (features) and the output (labels) for this problem.

2. Our goal is to use the images in the training data to learn a classifier that gives the label of a new handwritten digit. Suppose that a neural network has been trained for this task. How would you measure its accuracy and error rate?

3. In the python exercise, you divided the dataset to a training and test set. And then, you used a batch size of 32 for stochastic gradient descent. Write the steps of stochastic gradient descent with a batch size of 32. How many iterations would be in each epoch?

4. Suppose you were to use a logistic regression to learn a classifier that takes an input image and gives the label. What would be the number of parameters (weights and biases) you would have to learn?

5. Now, consider the case that the images get corrupted by noise and the pixel values get permuted. Which of the following approaches is more likely to suffer from accuracy and why? logistic regression, convolutional neural network.